(0x5f3759…

<u>Fast Inverse Square Root</u>

floating point $f_x = \overbrace{\underbrace{S_x}_{1\text{ bit}}\ \underbrace{E_x}_{b\text{ bits}}\ \underbrace{M_x}_{N-b-1\text{ bits}}}^{N\text{-bits}}$

sqrt only defined for [illegible], $S_x = 0$

$$\therefore f_x = \left(1 + \frac{M_x}{2^{N-b-1}}\right) 2^{E_x - (2^{b-1} - 1)}$$

For simplicity, let $L = 2^{N-b-1}$, which normalizes the mantissa M to $0 \leq M < 1$
& let $B = 2^{b-1} - 1$, which is the exponent bias.

$$f_x = \left(1 + \frac{M_x}{L}\right) 2^{E_x - B} \qquad (1)$$

Looking for $y = \frac{1}{\sqrt{x}} = x^{-\frac{1}{2}}$

Let f_x & f_y be floating [illegible] representations of x & y respectively.

$\therefore f_y = f_x^{-\frac{1}{2}}$ ignoring errors introduced by floating [illegible]

$$\log_2 f_y = \log_2 \left(f_x^{-\frac{1}{2}}\right)$$

$$\log_2 f_y = -\frac{1}{2} \log_2 f_x$$

$$\log_2 \left(\left(1 + \frac{M_y}{L}\right) 2^{E_y - B}\right) = -\frac{1}{2} \log_2 \left(\left(1 + \frac{M_x}{L}\right) 2^{E_x - B}\right)$$

$$\log_2 \left(1 + \frac{M_y}{L}\right) + \log_2 \left(2^{E_y - B}\right) = -\frac{1}{2} \left[\log_2 \left(1 + \frac{M_x}{L}\right) + \log_2 2^{E_x - B}\right]$$

$$\log_2 \left(1 + \frac{M_y}{L}\right) + E_y - B = -\frac{1}{2} \log_2 \left(1 + \frac{M_x}{L}\right) - \frac{1}{2} E_x + \frac{1}{2} B$$

$\log_2(1+$ [illegible] $) + E_y$ [illegible] $-\frac{1}{2}E_x +$ [illegible]

[illegible] $-\frac{1}{2}E_x -$ [illegible] (2)

[illegible] $1+$ [illegible] 1:

[illegible] $\log_2(1+m$ [illegible]

[illegible] $\log_2$ [illegible] θ_m [illegible]

[illegible]

substituting (3) into (2) we have:

$$\frac{M_y}{L} + \theta_y - E_y = -\frac{1}{2}\left(\frac{M_x}{L} + \theta_x\right) - \frac{1}{2}(E_x - 3\delta)$$

$$\frac{M_y}{L} + E_y \ \text{[illegible]}$$

$$E_y L + M_y = \text{[illegible]} - L\left(\tfrac{1}{2}\theta_x + \theta_y\right) - \text{[illegible]}$$

$$= -\tfrac{1}{2}M_x - L\left(\tfrac{1}{2}\theta_x + \theta_y\right) - \tfrac{1}{2}E_x \ \text{[illegible]}$$

$$E_y L + M_y = \tfrac{3}{2}\delta - (\text{[illegible]}) - \tfrac{1}{2}(E_x L + M_x)$$

Now [illegible] look at [illegible]

N-1 [illegible]

[illegible] 2^{N-1} [illegible] $= EL + M$

[illegible] if we take I_x and I_y [illegible] the [illegible] f_x & f_y, respectively, cast to integers, then we have

$$I_x = E_x L + M_x \quad , \quad I_y = E_y L + M_y .$$

So from (5), (6) [illegible]

$$I_y = L\left(\tfrac{3}{2}B - \left(\tfrac{1}{2}\sigma_x + \sigma_y\right)\right) - \tfrac{1}{2}I_x$$

or more simply

$$I_y = R - \tfrac{1}{2}I_x$$

where $R = L\left(\tfrac{3}{2}B - \left(\tfrac{1}{2}\sigma_x + \sigma_y\right)\right)$

And that's the basic technique: take floating point x [illegible] α [illegible] 1), & subtract our magic integer R [illegible] I_y. If you take [illegible] as a floating point you get y.

The error comes from the fact that σ_x & σ_y [illegible] specific values of x [illegible]

So we need to pick σ_x & σ_y as best as [illegible] an approximation for most [illegible] x.

For single precision [illegible]

$L = 2^{23}$

$B = 127$

$R = 2^{23}\left(\tfrac{3}{2}(127) - \left(\tfrac{1}{2}\sigma_x + \sigma_y\right)\right)$